A. Take home assignments, Sampling and estimation, winter semester 2012

Can be handed in personally to Daniel Thorburn at the lectures, by post or e-mail Daniel.Thorburn@stat.su.se , out in the departments mailbox (outside the elevators) or in his department post-box or handed to Nicklas Pettersson. The assignments must have reached DT or NP before 17.00.

The solutions must be your own. You are allowed to discuss the problems with each other on a superficial level but two solutions that resemble each other too much will get less credit (e.g. the same calculation errors or paragraphs with the same wordings). Note also that these assignments will influence the marks on the course. Some of them are difficult in order to be able to differentiate between you. Do not expect to be able to solve everything perfectly.

For Monday, January 30

1. For which of the following sampling problems can it be reasonable to assume a finite or infinite population. (Estimate population parameters or the latent model-distribution parameters). Motivate your answer with about 10 lines per question!

a) An auditor wants to find out the number of errors in the accounts of a savings bank. He selects 500 accounts out of 15 373 and writes to the account owner and asks whether the amount in the books are the correct value of their deposits in the bank at the end of the year. 479 of the account owners agree that this is the correct deposit. The remaining 21 are more closely examined and it turns out that in 12 cases the figures from the bank were correct but in 9 cases transactions were booked at the wrong year.

b) In a study of whether dyslexia increases in a Sweden 200 schools are asked about the number of pupils in the fifth grade who have a diagnosis of dyslexia (and the total number of pupils in that grade). In the study the schools are also asked about the measures they have taken to ensure that the pupils get the support they are entitled to. This study is repeated the following year.

c) A shop owner wants to know how their customers are treated in the shop and their opinion of the products for sale in the shop. A market institute hires two interviewers who stand outside the shop and interviews a random sample of the customers leaving the shop.

d) Statistics Sweden tries to estimate the gross national product of Sweden. A sample of enterprises are, in some detail, asked about their sales and their costs during January (and also about their production and changes in inventories).

e) 3245 randomly selected Swedes are asked about their attitude in a series of political issues. One wanted to know including their attitudes to the European Union, to the Euro, their political sympathies and their attitudes to nuclear power.

2. At a university a simple random sample of students are after the first test asked about how long much time they have allocated to the studies per week (non-anonymously and on the

web). When the sample is analysed it is divided into four groups. Did not partake in the test (N=48, n=12), did not pass (N=27, n=8), pass (N=52, n=14) and high pass (N=25, n=6).

Data not partake	8, 8, 24, 22, 7, 0, 0, 12, 7, 19, 2, 15
Not pass	27, 12, 8, 5, 17, 4, 25, 2
Pass	27, 21, 19, 15, 12, 28, 32, 20, 20, 20, 11, 17, 20, 14
High Pass	25, 20, 25, 25, 30, 10

Estimate the average number of hours allocated to the studies in the four groups, estimate the variances of the estimates and give uncertainty intervals.

Here you should view the problem as a finite sample problem. Discuss whether this is a reasonable assumption.

3. Today a common way of doing surveys is by using web-panels. A common way of doing such a survey is to first select a large random sample from the (Swedish) population. All of them are asked if they agree to be part of a web-panel and also asked to give their e-mail addresses. They are also asked a lot of background questions. This panel is then used as a frame in many studies in the future.

The real survey is done by selecting a random sample from the panel created above, sending out a questionnaire on the web. This gives a fast and cheap study since the answers go directly into the computer and are already checked and edited. You can get make a study in one week with a sample of 10000. The sample is in principle taken with probabilities inversely proportional to the tendency to agree to be in the panel in the first study. E.g. if 30 % of the women said yes and 15 % of the men, this is compensated by selecting men with twice the probability in the web survey. The web sample will thus have the correct proportions in all known background variables (e.g. age sex, region but also income, education, social group and political opinion (vote in last election) ...)

Discuss whether this is a recommendable procedure. Write an essay with a length corresponding to about one page in Word. Discussing the pros and cons?

4. Spooose that you are working as a consulting statistician. You are asked to construct an estimating plan for the following study. The client had had access to three lists on traffic accidents the road: Police reported accidents, patients hospitalized for traffic injuries and claims reported to insurance companies (with N1, N2, N3 cases on the three lists). One accident may be on none, one, two ort three of the lists. On the second and third list several cases may correspond to the same accident. The client had selected n1, n2 respectively n3 cases with SRS from the three lists and eliminated duplicates. The intended sampling frame was the merged list.

An incident that appears on multiple lists has thus a greater probability of selection than one who appears only once. The number of times it is recorded probably has some connection with the accident character. You realize that in order to obtain unbiased estimates you must

take the first and second order inclusion probabilities into account (and that they are functions of how many times the accident appears on the three lists).

Your task is to write a short PM or report to your principal pointing out the problems with varying inclusion probabilities, computing or suggesting ideas on how the inclusion probabilities can be computed, suggesting what questions to ask in order to be able to compute the probabilities.

To pass it is not required that you resolve all the parts - but you should justify this to the client and point out which problems remain.